Multiple-average-voice-based speech synthesis

MULTIPLE-AVERAGE-VOICE-BASED SPEECH SYNTHESIS

Pierre Lanchantin[†], *Mark J.F. Gales*[†], *Simon King*[*], *Junichi Yamagishi*[*]

[†]Cambridge University Engineering Department, Cambridge CB2 1PZ, UK
{pkl27,mjfg}@cam.ac.uk
[*]Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK
simon.king@ed.ac.uk, jyamagis@inf.ed.ac.uk

ABSTRACT

This paper describes a novel approach for the speaker adaptation of statistical parametric speech synthesis systems based on the interpolation of a set of average voice models (AVM). Recent results have shown that the quality/naturalness of adapted voices depends on the distance from the average voice model used for speaker adaptation. This suggests the use of several AVMs trained on carefully chosen speaker clusters from which a more suitable AVM can be selected/interpolated during the adaptation. In the proposed approach a set of AVMs, a multiple-AVM, is trained on distinct clusters of speakers which are iteratively re-assigned during the estimation process initialised according to metadata. During adaptation, each AVM from the multiple-AVM is first adapted towards the target speaker. The adapted means from the AVMs are then interpolated to yield the final speaker adapted mean for synthesis. It is shown, performing speaker adaptation on a corpus of British speakers with various regional accents, that the quality/naturalness of synthetic speech of adapted voices is significantly higher than when considering a single factor-independent AVM selected according to the target speaker characteristics.

Index Terms— HMM-Based speech synthesis, speaker adaptation, multiple average voice model, cluster adaptive training

1. INTRODUCTION

Statistical parametric speech synthesis based on hidden Markov models (HMMs) is now a well-established approach which can generate natural-sounding synthetic speech [1]. It has several advantages compared to concatenative speech synthesis [2] such as small footprint [3, 4, 5], robustness to non-ideal speech [6], but also flexibility to change the voice characteristics [7, 8, 9, 10, 11]. Adaptation techniques - initially derived from the speech recognition field - can be applied giving the ability to create new voices using only a small amount of adaptation data from a target speaker. In the average-voice-based speech synthesis framework [6] - referred as the AVM framework in this work - an average voice model (AVM) is used as the seed of the adaptation process. Its characteristics directly affects the ones of the speech generated from the resulting adapted model. Recent analyses of speaker adaptation performance [6, 12] have shown that the quality/naturalness of adapted voices is moderately correlated with the distance between the AVM and the target voice; transforming the AVM towards a distant speaker tends to degrade the synthesised speech quality. It was then found in [13] that AVMs trained on perceptually similar speakers provide better performance than global models, even through the latter are trained on more data. In other words, it is better to train AVMs on a smaller number of carefully selected speakers than a large number of arbitrary speakers. These results suggest that adaptive HMM-based speech synthesis systems can take advantage of using several AVMs from which the most appropriate AVM can be selected or even combined. Hence, in [14], AVM mean vectors were combined by summing the linearly transformed mean vector of output distributions of each model, the contributing rate of each model being estimated by extended speaker adaptive training (ESAT [15]). An other possibility, which will be presented in this paper, is to interpolate the AVM mean vectors in the same way than in the cluster adaptive training (CAT) framework. This has the advantage to consider, for each stream, the set of decision trees of all the AVMs. It should then allow a wide variety of possible contexts to be produced as there is an intersect of context trees.

In the light of these points, this paper examines a novel approach for speaker adaptation, based on a *Multiple-AVM*, which can be seen as an hybrid between the AVM and the CAT approaches. In the same fashion than CAT, the set of AVM mean vectors constitutes an eigenspace in which the adapted mean vector is interpolated. However, in contrast with CAT, each AVM is first adapted towards the speaker using constrained structural maximum a posteriori linear regression (CSMAPLR [6]), which suggests a better tuning to the target speaker of the eigenspace in which the interpolation takes place. Each AVM is trained separately on clusters of carefully selected speakers, re-assigned at each iteration. Any commonalities across speakers are not exploited, in contrast with CAT. However, this considerably simplifies the training process, especially when the amount of training data and number of clusters gets large. It is shown via experiments that the performance of speaker adaptation can be significantly improved in terms of quality/naturalness using the proposed approach, compared to a conventional one using a single AVM selected according to the target speaker characteristics.

The rest of paper is laid out as follows. Section 2 describes the proposed approach in contrast with CAT. The method is perceptually evaluated in Section 3 and section 4 concludes.

2. PROPOSED APPROACH

2.1. Cluster Adaptive Training

The proposed approach being closely similar to CAT, we briefly describe its framework. CAT was initially proposed for fast adaptation in speech recognition [16] and recently extended to speech synthesis for polyglot text-to-speech [17], combination of multiple high quality corpora [18] and for the control of specific factors of the generated voice in [19]. In the speech synthesis extension, the structure of the CAT model includes multiple clusters, each with their own decision trees. A bias cluster is considered containing covari-

This research was supported by ESPRC Programme Grant, grant no. EP/I031022/1 (Natural Speech Technology)

ances and mixture weight parameters while other clusters contain only mean vectors. The emission probability of an observation vector $\boldsymbol{o}_t$ at frame t given Gaussian component m, speaker s, and a set of model parameters θ can be expressed as

$$p(\boldsymbol{o}_t|m,s;\theta) = \mathcal{N}\left(\boldsymbol{o}_t;\ \boldsymbol{\mu}_m^{(s)},\ \boldsymbol{\Sigma}_m^{(s)}\right) \tag{1}$$

$\boldsymbol{\mu}_m^{(s)}$ and $\boldsymbol{\Sigma}_m^{(s)}$ are the interpolated mean vector and the covariance matrix of component m, respectively. The latter is shared over all the clusters via the bias cluster. The set of P clusters defines an eigenspace representing all possible speakers in which the position of a speaker s is given by a vector of CAT interpolation weights

$$\boldsymbol{\lambda}_{q(m)}^{(s)} = \left[1\ \lambda_{2,q(m)}^{(s)}\ \ldots\ \lambda_{P,q(m)}^{(s)}\right]^\top \tag{2}$$

where each $\lambda_{p,q(m)}^{(s)}$ is the CAT interpolation weight for cluster p associated with weight $q(m) \in \mathcal{Q}$ of the component m, $\mathcal{Q}$ being the set of Q disjoint cluster weight classes. The first weight is equal to 1 as the first cluster is specified as a bias one. Note that HMM-based speech synthesis systems making use of multiple streams, each stream have its own eigenspace. The mean vector of the Gaussian component $\boldsymbol{\mu}_m^{(s)}$ is found by linearly combining the mean vectors of each cluster according to the vector of interpolation weights, as

$$\boldsymbol{\mu}_m^{(s)} = \boldsymbol{M}_m \boldsymbol{\lambda}_{q(m)}^{(s)} \tag{3}$$

where $\boldsymbol{M}_m$ is the matrix of P cluster class mean vectors $\boldsymbol{\mu}_{l(p,m)}$ for a component m, as $\boldsymbol{M}_m = [\boldsymbol{\mu}_{l(1,m)}\ \ldots\ \boldsymbol{\mu}_{l(P,m)}]$ where $l(p,m)$ is the leaf node for component m in decision trees of AVM p.

The parameters are estimated using an expectation-maximisation algorithm in which the canonical parameters, the CAT weights and the decision trees are each updated separately in a similar way than speaker adaptive training (SAT [20],[21]). Though this joint modelling is optimal in the sense that it allows systems to share parameters and data to not be fragmented, it can however become computationally very expensive as the amount of training data and number of clusters gets large.

2.2. Multiple-AVM

In the proposed approach, the structure of the model is slightly different. Each cluster is an AVM trained independently, having its own decision trees, and containing mean vectors but also covariances and mixture weights parameters, so that no bias cluster is considered. This set of AVMs will be referred as *Multiple-AVM* in the following description. During adaptation, in the same fashion as CAT, the matrix composed of the mean vectors of each AVM for a given component m defines an "eigenspace"[1] in which the mean vector m for the target speaker s is simply given by an interpolation weight vector, $\boldsymbol{\lambda}_{q(m)}^{(s)}$

$$\boldsymbol{\lambda}_{q(m)}^{(s)} = \left[\lambda_{1,q(m)}^{(s)}\ \ldots\ \lambda_{P,q(m)}^{(s)}\right]^\top \tag{4}$$

However, a major difference is that this eigenspace is first adapted towards the target speaker before the interpolation so that the mean matrix now depends on the target speaker s, as

$$\boldsymbol{\mu}_m^{(s)} = \boldsymbol{M}_m^{(s)} \boldsymbol{\lambda}_{q(m)}^{(s)} \tag{5}$$

where $\boldsymbol{M}_m^{(s)}$ is the matrix of P AVM mean vectors $\boldsymbol{\mu}_{l(p,m)}^{(s)}$ for a component m, as $\boldsymbol{M}_m^{(s)} = [\boldsymbol{\mu}_{l(1,m)}^{(s)}\ \ldots\ \boldsymbol{\mu}_{l(P,m)}^{(s)}]$. This should enable a better tuning to the individual speaker of the eigenspace in which the interpolation takes place, the AVM usually giving better similarity performance than CAT when the amount of adaptation data is adequate. An important consideration for the multiple AVM model is that the *space* in which the means are interpolated needs to be consistent. There are three distinct spaces[2] here: the original space, the primary AVM space and the secondary AVM space. Note in the primary and secondary AVM space, the covariance matrices are diagonal whereas they are full in the original space.

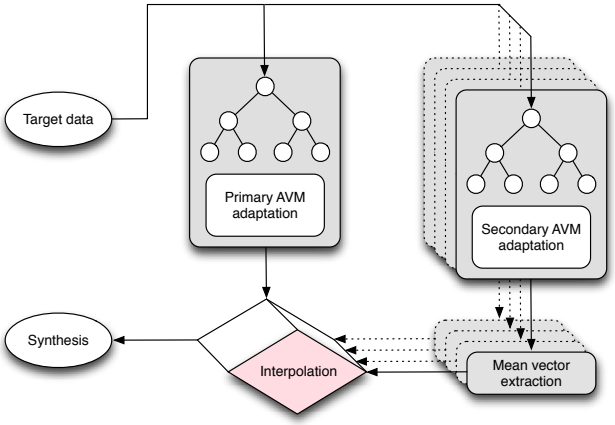

Fig. 1. The Multiple-AVM framework.

2.2.1. *Multiple-AVM Adaptation*

We now describe the adaptation procedure of the multiple-AVM, illustrated in Fig. 1. The emission probability of an observation vector $\boldsymbol{o}_t$ at frame t given component m, AVM p, speaker s, and a set of model parameters θ can be expressed as

$$p(\boldsymbol{o}_t|m,p,s;\theta) = \mathcal{N}\left(\boldsymbol{o}_t;\ \boldsymbol{\mu}_{l(p,m)}^{(s)},\ \boldsymbol{\Sigma}_{l(p,m)}^{(s)}\right) \tag{6}$$

Primary AVM We first consider, as in a conventional AVM-based system, an unique AVM selected from a set of several AVMs for the constrained structural maximum a posteriori linear regression (CSMAPLR [6]) adaptation towards a target speaker s. The mean of the Gaussian component m defined in (6) of AVM of index 1 is given by

$$\boldsymbol{\mu}_{l(1,m)}^{(s)} = \dot{\boldsymbol{A}}_{1,r(m)}^{(s)} \boldsymbol{\mu}_{l(1,m)} + \dot{\boldsymbol{b}}_{1,r(m)}^{(s)} \tag{7}$$

where $\boldsymbol{\mu}_{l(1,m)}$, $\dot{\boldsymbol{A}}_{1,r(m)}^{(s)}$ and $\dot{\boldsymbol{b}}_{1,r(m)}^{(s)}$ are the mean vector of component m, the CSMAPLR linear transformation matrix and bias vector[3] for speaker s associated with regression class $r(m)$, for AVM 1, respectively. This AVM is selected from a set of AVMs according to the likelihood of the adaptation data given the models, the selected one being the one maximising this value. We will refer to it as the *primary* AVM .

Covariance matrix In the proposed approach, considering that no bias cluster is used, the covariance of the component m for the

[1] no orthogonality constraints are considered here.

[2] Strictly every CMLLR/CSMAPLR transform defines a space. For simplicity rather than considering all these spaces, the space for each component is considered.

[3] In the following, for brevity in notation, we will omit to indicate bias in transformations, which however must be taken into account.

adapted speaker must be defined. We assume that the covariance $\boldsymbol{\Sigma}_m^{(s)}$ of a component for the adapted speaker is given by the adapted primary AVM so as

$$\boldsymbol{\Sigma}_m^{(s)} = \boldsymbol{\Sigma}_{l(1,m)}^{(s)} = \dot{\boldsymbol{A}}_{1,r(m)}^{(s)} \boldsymbol{\Sigma}_{l(1,m)} \dot{\boldsymbol{A}}_{1,r(m)}^{(s)\top} \tag{8}$$

where $\boldsymbol{\Sigma}_{l(1,m)}$ is the covariance matrix of component m for AVM 1.

Secondary AVMs As the covariance matrix of the primary AVM is used during the interpolation, we will first express the mean of the secondary AVMs in the primary AVM space, since this will allow diagonal covariance matrix MLLR to be used. To do so, we first express the secondary AVMs mean in the original space by applying the CSMAPLR transform $\dot{\boldsymbol{A}}_{p,r(m)}^{(s)}$ for speaker s associated with regression class $r(m)$, for AVM p. Then the inverse primary CSMAPLR transform $\dot{\boldsymbol{A}}_{1,r(m)}^{(s)-1}$ is applied to yield a mean in the primary space. As CSMAPLR (and CMLLR) transforms simultaneously adapt both the means and variances, the adapted primary AVM means are expected to be better matched than the secondary AVM means in the primary space. To address this a MLLR transform $\hat{\boldsymbol{A}}_{p,r(m)}^{(s)}$ is estimated on the transformed mean (this is applied to both the primary and secondary AVM means). In this work, we will approximate this whole transformation by a MLLR transform as[4]

$$\boldsymbol{A}_{p,r(m)}^{(s)} \approx \hat{\boldsymbol{A}}_{p,r(m)}^{(s)} \dot{\boldsymbol{A}}_{1,r(m)}^{(s)-1} \dot{\boldsymbol{A}}_{p,r(m)}^{(s)} \tag{9}$$

$\boldsymbol{A}_{p,r(m)}^{(s)}$ is the MLLR mean linear transformation matrix for speaker s associated with regression class $r(m)$, for AVM p. For consistency, the MLLR transform is also estimated for the primary AVM. The interpolation is done in the original space[5] so that the transformed mean for all AVMs is given by

$$\boldsymbol{\mu}_{l(p,m)}^{(s)} = \dot{\boldsymbol{A}}_{1,r(m)}^{(s)} \boldsymbol{A}_{p,r(m)}^{(s)} \boldsymbol{\mu}_{l(p,m)} \tag{10}$$

2.2.2. *Estimation of the interpolation weights*

The vector of interpolation weight $\boldsymbol{\lambda}_q^{(s)}$ is estimated by maximum likelihood in the same way than in [16] for each AVM weight class $q \in \mathcal{Q}$, but considering the adapted mean matrix $\boldsymbol{M}_m^{(s)}$ towards the speaker s so as

$$\boldsymbol{\lambda}_q^{(s)} = \boldsymbol{G}_q^{(s)-1} \boldsymbol{k}_q^{(s)} \tag{11}$$

where the accumulated statistics $\boldsymbol{G}_q^{(s)}$ and $\boldsymbol{k}_q^{(s)}$ are given by

$$\boldsymbol{G}_q^{(s)} = \sum_{m \in q} \boldsymbol{M}_m^{(s)\top} \boldsymbol{\Sigma}_m^{(s)-1} \boldsymbol{M}_m^{(s)} \sum_t \gamma_m^{(s)}(t) \tag{12}$$

$$\boldsymbol{k}_q^{(s)} = \sum_{m \in q} \boldsymbol{M}_m^{(s)\top} \boldsymbol{\Sigma}_m^{(s)-1} \sum_t \gamma_m^{(s)}(t) \boldsymbol{o}(t) \tag{13}$$

where $\gamma_m^{(s)}(t)$ is the occupancy probability of component m for speaker s at time t.

[4]It would have been better to apply SMAPLR to be consistent with the fact that we used a CSMAPLR. The exact form of (9) will be examined in future work.

[5]Regression classes for CSMAPLR are determined according to the primary AVM decision tree. The linear transforms must also be applied to secondary AVMs for which components were tied according to different decision trees. In order to avoid mismatches, a simple solution is to untie the model set used for the adaptation (the number of models used during the adaptation being relatively small).

2.2.3. *Multiple-AVM Training*

Each AVM is trained separately on its own portion of data. Any commonalities across speakers are not exploited, in contrast with CAT. However this considerably simplifies the training process, especially when the amount of training data and number of clusters gets large. The partitioning of the training data should be done carefully as this determined the eigenspace in which the adaptation is performed. One way to do the partitioning is to first cluster speakers according to discriminating factors such as the gender, the age or the regional accent of speakers. Selection can then be done according to metadata associated to the speaker database for different values or range of values of the selected factors. However metadata is potentially unreliable. To address this a re-assignment of the speakers according to the likelihood of the speakers data given each model can be performed during. The process is:

1. *Initialisation*: initial speaker clusters are built according to metadata;
2. *Multiple-AVM Training*: AVMs are trained for each cluster using SAT;
3. *Speaker re-assignment*: speakers are re-assigned to clusters according to the likelihood of speakers data given each model

Steps 2 and 3 are repeated until the speaker assignment of each cluster remains stable or when a fixed number of iterations is reached. Note that after the training, the metadata associated initially to the clusters may become irrelevant due to the different speaker repartition among the clusters.

3. EXPERIMENTS

In order to valid our approach, we ran an experiment based on a corpus of British speakers with different regional accents, the latter being the discriminating factor. We wanted to assess the improvement in terms of quality/naturalness of the proposed approach compared to the one based on a single AVM selected from a set of AVMs. Similarity to the target speaker was not evaluated during these experiment as the differences between the baseline and multiple-AVM approaches were not judge significantly different in an initial informal listening test. The topology of the models was similar to the one used for the Nitech-HTS 2005 system ([22]). Speech data was sampled at 48 kHz. Each observation vector consisted of 60 Mel-cepstral coefficients [23], logarithmic fundamental frequency (log F_0) values, 25-band aperiodicities, and their first and second derivatives ($3 \times (60 + 25 + 1) = 256$) extracted every 5ms. Five-state, left-to-right, no-skip hidden semi-Markov models (HSMMs [24]) were used. A multi-space probability distribution (MSD) [25] was used to model $logF_0$ sequences consisting of voiced and unvoiced observations.

3.1. British Multiple-AVM Training

Two AVMs were considered as components of the multiple-AVM. They were trained using speaker re-assignment after initialisation using the metadata as presented in section 2.2.3. One of the AVM was initialised on a selection of 106 English speakers, the other being initialised on a selection of 181 Scottish speakers. Only a few speakers were re-assigned, even after the first iteration. In contrast, a similar experiment ran with age and gender as discriminative factors had shown a large amount of re-assigned speakers during the first iterations of the training process. This suggests that accent is a "hard" boundary, well defined by the metadata, compared to gender and age for which metadata may be unreliable. As only a few

speakers were re-assigned, the AVMs will be denoted as English and Scottish AVM for convenience. However, for other discriminative factors (e.g age or gender), this might not be the case as the final assignment of speakers can differ strongly from the initial assignment, the initial metadata value used for the selection becoming irrelevant. Note that the main objective of the training procedure is to build a set of AVMs that best represent the speaker space since it will be used as a basis for the interpolation, the meaning of each AVM is not important. Training data consisted of 57018 utterances for the English AVM and 72708 utterances for the Scottish one. 2926 questions were used for the decision tree-based context clustering for both AVMs. The sizes of decision trees were controlled by changing the scaling factor α for the model complexity penalty term of the minimum description length (MDL) criterion [27]. When $\alpha = 1$, the number of leaf nodes for Mel-cepstrum, log F_0 and band aperiodicities were 10153, 111358 and 4512 respectively for the English AVM (5334102 parameters in total) and 9271, 71995 and 4208 respectively for the Scottish AVM (4616715 parameters in total) .

3.2. British Multiple-AVM Adaptation

Four target speakers with different British accents were selected: a 22 years old male English speaker from Surrey, a 51 years old female Scottish/English mixed speaker from Ayrshire, a 42 years old male Scottish speaker from East Lothian, and a 66 years old female Scottish speaker from Glasgow, respectively referred as speaker A, B, C, and D in the following description. The set $\mathcal{Q}$ includes weight classes assigned to each stream - mel-cepstral coefficients (mcep), logarithmic fundamental frequency (lf_0) and its first (dlf_0) and second derivative (ddlf_0), and band aperiodicities (bap) - and to the duration of each of the 5 states of the HSMM $(\text{d}_1, \ldots, \text{d}_5)$ for both AVMs , which represents a total of 20 weights to be estimated during the AVMs' interpolation. 10 sentences ($\sim$30s) were used for the adaptation of the multiple-AVM. The English model was selected as the primary AVM for speaker A whereas the Scottish model was selected for speaker B,C and D. Both primary and secondary model were adapted towards the considered target speaker before interpolation, according to the method presented in section 2.2.1. The estimated weight vectors for each speaker are presented in Table 3.2. As expected the majority of the weight is assigned to the primary AVM. However for some of the streams of speakers A and D, more weight is assigned to the secondary AVM than the primary one. This indicates that it may be beneficial to perform speaker re-assignment at the stream level during the training of the AVMs.

Table 1. Estimated interpolation weights for each target speaker, AVM and stream.

		Compound					Duration				
Spk	AVM	mcep	lf_0	dlf_0	ddlf_0	bap	d_1	d_2	d_3	d_4	d_5
A	Sco	0.45	0.63	0.10	0.43	0.40	-0.02	0.03	0.01	0.27	0.21
	Eng	0.58	0.37	0.92	0.70	0.61	1.01	0.88	1.02	0.57	0.69
B	Sco	0.58	0.61	0.56	0.82	0.73	0.83	0.81	0.78	0.99	0.94
	Eng	0.44	0.39	0.51	0.25	0.27	0.13	0.15	0.25	0.06	0.07
C	Sco	0.69	0.85	0.72	0.78	0.70	0.70	0.89	0.91	1.00	0.92
	Eng	0.36	0.15	0.31	0.28	0.30	0.30	0.09	0.09	0.00	0.09
D	Sco	0.68	0.81	0.05	0.00	0.73	1.16	0.87	0.84	1.07	1.04
	Eng	0.36	0.19	0.96	0.97	0.27	-0.14	0.12	0.17	-0.03	-0.05

3.3. Subjective evaluation

A subjective preference listening test using controlled perception experiment booths was conducted on 30 listeners. For each of the 4 target speakers, 20 sentences were selected randomly from 4 different test corpora with different genres[6]. Each of the 80 sentences was synthesised[7] considering 2 adapted AVMs: the one closest to the target (selection approach) and the multiple-AVM described in this paper. Each pair was presented in random order to each listener. For a given pair, listeners were asked to choose the one of the two samples they preferred. In the case of no preference, they were asked to randomly select one of the two proposed samples. As mentioned earlier, the proximity to the target speaker was not evaluated as it was not found significantly different during an informal listening test. The samples generated by the multiple-AVM were significantly preferred to the conventional AVM one with a preference score of 57.1% (p-value$<10^{-4}$) as presented on Fig. 2. Speaker C benefited the most of the proposed approach with a preference score of 58.4% (p-value$<10^{-4}$) followed by speaker A, and D with preference score of 58.0% (p-value$<10^{-2}$) and 56.3% (p-value$<10^{-3}$), respectively. Finally, speaker B had a lower, but still significant, preference score of 55.6% (p-value$<5.10^{-2}$). These results indicate that the proposed framework has the potential to improve the quality of the synthesised speech compared to the selection approach. We expect that the wider variety of possible contexts which can be produced by the intersect of context trees contributed to the preference.

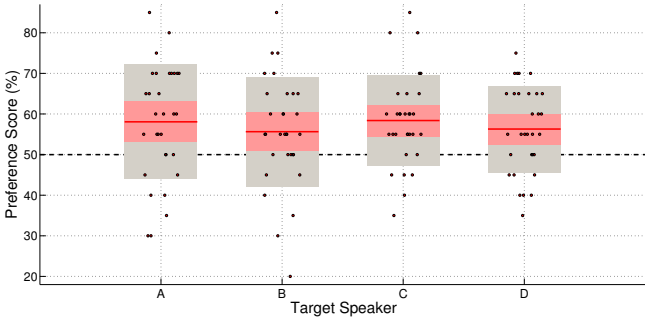

Fig. 2. Results of the listening test. Points representing average listeners preference score for each speaker (A, B, C, and D) are layed over a 1.96 standard error of the mean (SER) (95% confidence interval) in pink and a 1 standard deviation (SD) in grey.

4. CONCLUSION

This paper examined a novel approach for speaker adaptation based on multiple-AVM. As in the CAT framework, the set of decision trees of all the AVMs is considered during the adaption. However, the training stage is computationally less expensive than CAT, as the amount of training data and clusters gets larger. Furthermore, the whole set of AVM is first adapted towards the speaker before the interpolation stage which suggests a better tuning to the individual speaker of the space in which the interpolation takes place. Experiments ran on a corpus of British speakers with various regional accents confirmed a significant preference for samples generated from the adapted multiple-AVM compared to an AVM selection approach. Future work includes the run of larger experiments considering different factors such as age or gender involving a larger number of AVMs, and the comparison with CAT based adaptation. On a longer term, we plan to improve the training of multiple-AVM by performing speaker re-assignment at a the stream level and to develop an adaptive training approach in the proposed framework.

Acknowledgments The authors would like to thank C. Veaux for the selection of British speakers and V. Karaiskos for his contribution to the listening test.

[6] broadcast, news, novel and semantically unpredictable sentences.

[7] During speech generation, acoustic feature parameters were generated from the adapted MSD-HSMMs considering global variance [26].

5. REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996, pp. 373–376.

[3] Y. Morioka, S. Kataoka, H. Zen, Y. Nankaku, K. Tokuda, and T. Kitamura, "Miniaturization of HMM-based speech synthesis," in *Proc. Autumn Meeting of ASJ*, 2004, pp. 325–326.

[4] S.-J. Kim, J.-J. Kim, and M.-S. Hahn, "Hmm-based Korean speech synthesis system for hand-held devices," *IEEE Trans. Consum. Electron.*, vol. 52, no. 4, pp. 13841390, 2006.

[5] A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, "Quantized HMMs for low footprint text-to-speech synthesis," in *Proc. Interspeech*, 2010, p. 837840.

[6] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, January 2009.

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Proc. Eurospeech*, 1997, p. 25232526.

[8] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP*, 2001, pp. 805–808.

[9] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. ICSLP*, 2002, pp. 1269–1272.

[10] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[11] N. Obin, P. Lanchantin, A. Lacheret-Dujour, and X. Rodet, "Discrete/continuous modelling of speaking style in HMM-based speech synthesis: design and evaluation," in *Proc. Interspeech*, 2011, pp. 2785–2788.

[12] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y. J. Wu, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis-analysis and application of TTS system built on various ASR corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, July 2010.

[13] R. Dall, C. Veaux, J. Yamagishi, and S. King, "Analysis of speaker clustering strategies for HMM-based speech synthesis," in *Proc. Interspeech*, 2012.

[14] J. Isogai, J. Yamagishi, and T. Kobayashi, "Model adaptation and adaptive training using esat algorithm for hmm-based speech synthesis," in *Proc. Interspeech*, 2005, pp. 2597–2600.

[15] M. J. F. Gales, "Multiple-cluster adaptive training schemes," in *Proc. Icassp 2001*, 2001, pp. 361–364.

[16] M.J.F. Gales, "Cluster adaptive training of hidden markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 8, pp. 417–428, 2000.

[17] H. Zen, N. Braunschweiler, S. Buchholz, M. J. F. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, August 2012.

[18] V. Wan, J. Latorre, K. K. Chin, L. Chen, M. J. F. Gales, H. Zen, K. M. Knill, and M. Akamine, "Combining multiple high quality corpora for improving HMM-TTS," in *Proc. Interspeech*, 2012.

[19] J. Latorre, V. Wan, M. J. F. Gales, L. Chen, K. K. Chin, K. M. Knill, and M. Akamine, "Speech factorization for HMM-TTS based on cluster adaptive training," in *Proc. Interspeech*, 2012.

[20] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization," in *Proc. ICASSP*, 1997, vol. 2, pp. 1043–1046.

[21] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[22] H. Zen, T. Toda, M. Nakamura, and T. Tokuda, "Details of the nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325333, 2007.

[23] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, p. 137140.

[24] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825834, 2007.

[25] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synthesis Workshop*, 2002.

[26] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816824, 2007.

[27] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. Eurospeech*, 1997, p. 99102.